Cours 3
Arbres de décision
Classification bayésienne
Estimation des densités de probabilité

Catherine ACHARD
Institut des Systèmes Intelligents et de Robotique

catherine.achard@sorbonne-universite.fr

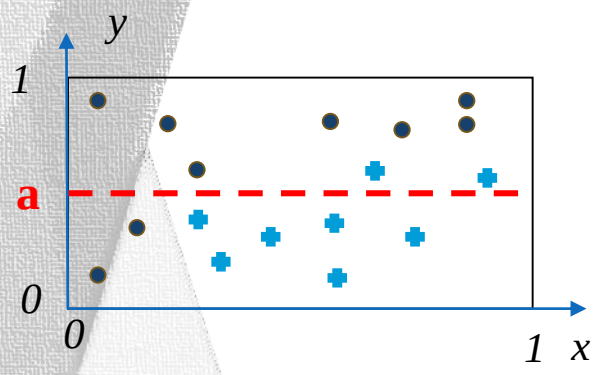
Méthode discriminative
arbre de décision

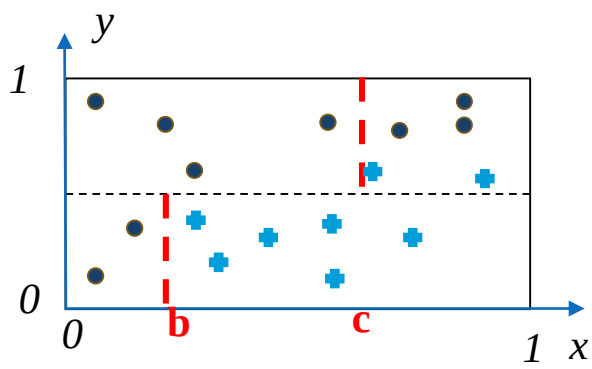
Arbre de décision

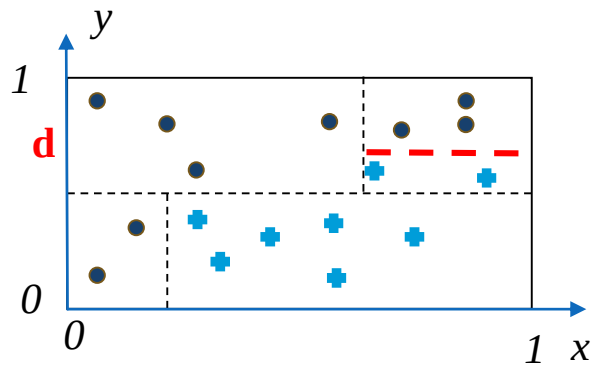
C'est une méthode discriminative

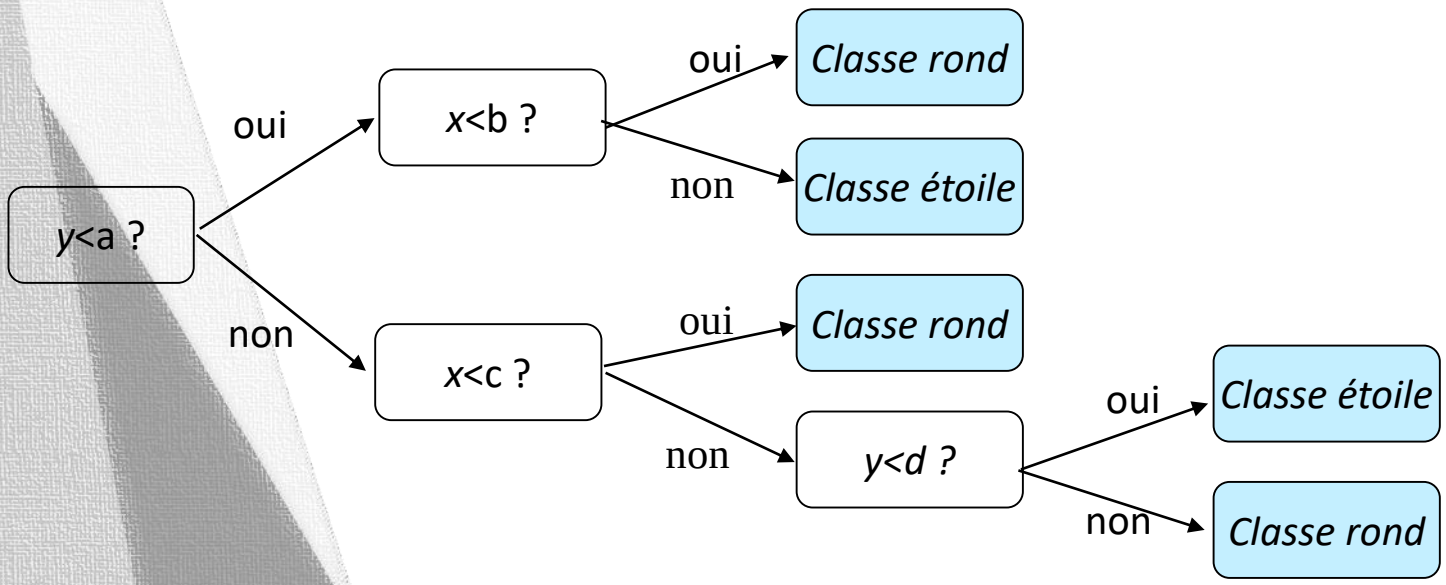
Idée

- Classer avec un ensemble de règles.
- Une suite de décisions permet de partitionner l'espace en régions homogènes
- La difficulté consiste à créer l'arbre à partir de la base d'exemple étiquetée









Arbre de décision

But

Trouver l'ordre le plus cohérent des questions qui amènera le plus rapidement à la solution

Initialisation

tous les exemples sont dans le même nœud no

Procédure construit_arbre(no)

- Si no est une feuille
  - Affecter une classe à no
- Sinon
  - Choisir la meilleure question et partitionner no en no_1 et no_2
  - Construit_arbre(no_1)
  - Construit_arbre(no_2)
- Fin si

Problème 1 : comment le savoir

Problème 2 : quelle classe ?

Problème 3 : comment choisir la meilleure question ?

Toutes les difficultés résident dans la réponse à ces trois problèmes

L'algorithme est très général. Plusieurs solutions proposées en fonction de la réponse à ces problèmes

Arbre de décision

Problème 1: comment décider qu'un nœud est une feuille

- Quand tous les exemples du nœud appartiennent à la même classe
- Quand tous les exemples du nœud ont le même vecteur de paramètres
- Quand le nombre d'exemples du nœud est inférieur à un seuil
- Quand une classe est largement majoritaire dans le nœud
- En contrôlant les performances en généralisation sur une base de validation

Problème 2: Quelle classe attribuer à une feuille

On affecte au nœud la classe majoritaire de ses exemples

Problème 3: Comment choisir la meilleure question

Plusieurs méthodes. Exemple : on utilise la **théorie de l'information**.

Entropie sur X conditionnée par q (quantité d'information qu'il reste sur X quand on connait la réponse à la question q)

$$H(X/q) = -\sum_{u,v} p(X = u, q = v) log_2(X = u/q = v)$$

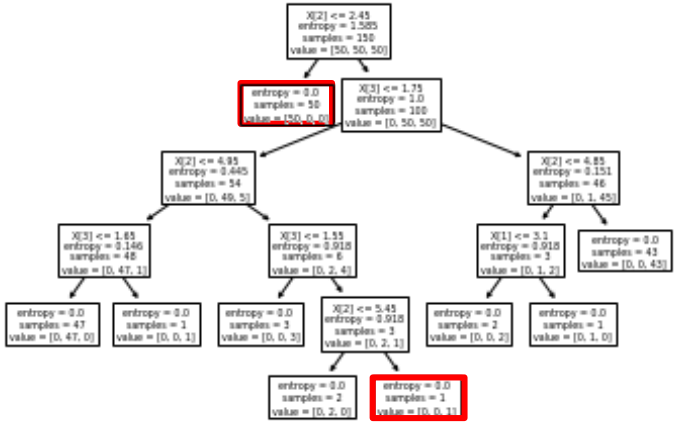
$H(X/q)$: quantité d'information contenue dans X si on connaît q.

On va rechercher la question q qui **minimise** cette quantité d'information restante (on voudrait que q nous amène toutes les connaissances)

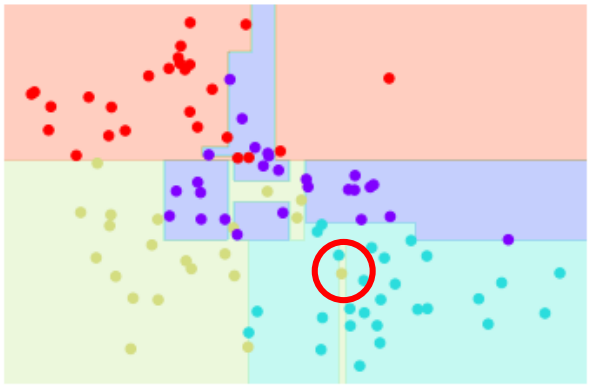
Arbre de décision

Problème

- Les arbres peuvent ne pas être équilibrés ➔ temps de parcours dépendant des exemples



- Présence de branches longues, peu représentatives qui nuisent à la généralisation générées par des exemples atypiques. Il faut alors les élaguer ou mettre des critères sur la profondeur, la taille des feuilles,… ou utiliser une base de validation
- Les arbres sont instables ➔ très forte dépendance aux données ➔ variance élevée



Solution:

➔**Construction d'une forêt d'arbres aléatoires grâce au bagging**

Arbre de décision

Bagging

But

- Réduire la variance des prédictions

Principe

- Construire aléatoirement plusieurs sous-ensembles d'apprentissage par tirage avec remise. Chaque sous-ensemble est appelé bootstrap
- Apprendre un arbre sur chaque sous-ensemble
- Fusionner les résultats des classifieurs

Random forest : arbre décisionnel + bagging

- Construire aléatoirement plusieurs sous-ensembles d'apprentissage
- Construire un arbre sur chaque sous-ensemble : si les données sont de dimension n, à chaque nœud, tirer aléatoirement $n' < n$ dimensions pour construire l'arbre. Ceci amène à des arbres moins corrélés
- Fusionner les résultats des arbres par vote majoritaire

Arbre de décision

Déterminer l'arbre de décision avec les exemples ci-dessous:

	Ciel	Température	Humidité	Vent	Décision
	q1	q2	q3	q4	
ex1	soleil	Chaud	Normale	oui	randonnée
ex2	Nuage	Froid	Haute	non	randonnée
ex3	soleil	Froid	Normale	oui	randonnée
ex4	soleil	Chaud	Normale	non	randonnée
ex5	Nuage	Froid	Normale	non	Pas randonnée
ex6	Nuage	Froid	Haute	oui	Pas randonnée
ex7	soleil	Chaud	Haute	non	Pas randonnée
ex8	soleil	Froid	Haute	oui	Pas randonnée

Arbre de décision

Il faut estimer l'entropie conditionnée par chaque question

$$H(X/q) = -\sum_{u,v} \mathrm{p}(\mathrm{X} = \mathrm{u}, \mathrm{q} = \mathrm{v}) log_2(p(X = u/q = v))$$

Et donc

- $H(X/q1)$
- $H(X/q2)$
- $H(X/q3)$
- $H(X/q4)$

	Ciel	Température	Humidité	Vent	Décision
	q1	q2	q3	q4	
ex1	soleil	Chaud	Normale	oui	randonnée
ex2	Nuage	Froid	Haute	non	randonnée
ex3	soleil	Froid	Normale	oui	randonnée
ex4	soleil	Chaud	Normale	non	randonnée
ex5	Nuage	Froid	Normale	non	Pas randonnée
ex6	Nuage	Froid	Haute	oui	Pas randonnée
ex7	soleil	Chaud	Haute	non	Pas randonnée
ex8	soleil	Froid	Haute	oui	Pas randonnée

Commençons par $q1$:

$$H(X/q1) = \begin{array}{c} -\mathrm{p}(\mathrm{rand}, \mathrm{q1} = \mathrm{soleil})\, log_2(p(\mathrm{rand}/q1 = \mathrm{soleil})) \\ -\mathrm{p}(\mathrm{pasrand}, \mathrm{q1} = \mathrm{soleil})\, log_2(p(\mathrm{pasrand}/q1 = \mathrm{soleil})) \\ -\mathrm{p}(\mathrm{rand}, \mathrm{q1} = \mathrm{nuage})\, log_2(p(\mathrm{rand}/q1 = \mathrm{nuage})) \\ -\mathrm{p}(\mathrm{pasrand}, \mathrm{q1} = \mathrm{nuage})\, log_2(p(\mathrm{pasrand}/q1 = \mathrm{nuage})) \end{array}$$

Arbre de décision

Il faut estimer l'entropie conditionnée par chaque question

$$H(X/q1) = -\sum_{u,v} \mathrm{p}(\mathrm{X} = \mathrm{u}, \mathrm{q1} = \mathrm{v}) log_2(p(X = u/q1 = v))$$

Pour la question q1

Ciel	soleil	nuage
rand	p(rand,soleil)	p(rand,nuage)
Pas rand	p(pas rand,soleil)	p(pas rand,nuage)

Arbre de décision

Il faut estimer l'entropie conditionnée par chaque question

$$H(X/q1) = -\sum_{u,v} \mathrm{p}(\mathrm{X} = \mathrm{u}, \mathrm{q1} = \mathrm{v}) log_2(p(X = u/q1 = v))$$

Pour la question q1

Ciel	soleil	nuage
rand	3	1
Pas rand	2	2
	5	3

	Ciel	Température	Humidité	Vent	Décision
	q1	q2	q3	q4	
ex1	soleil	Chaud	Normale	oui	randonnée
ex2	Nuage	Froid	Haute	non	randonnée
ex3	soleil	Froid	Normale	oui	randonnée
ex4	soleil	Chaud	Normale	non	randonnée
ex5	Nuage	Froid	Normale	non	Pas randonnée
ex6	Nuage	Froid	Haute	oui	Pas randonnée
ex7	soleil	Chaud	Haute	non	Pas randonnée
ex8	soleil	Froid	Haute	oui	Pas randonnée

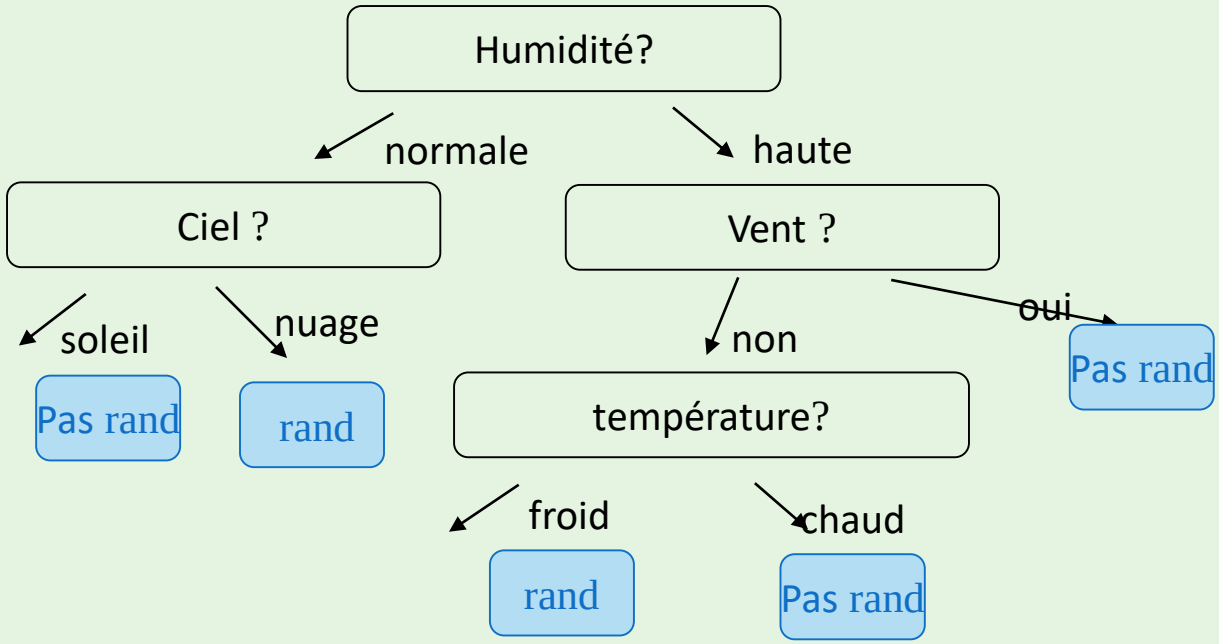
$$H(X/q1) = \begin{matrix} -\mathrm{p}(\mathrm{rand}, \mathrm{q1} = \mathrm{soleil})\ log_2(p(\mathrm{rand}/q1 = \mathrm{soleil})) \\ -\mathrm{p}(\mathrm{pasrand}, \mathrm{q1} = \mathrm{soleil})\ log_2(p(\mathrm{pasrand}/q1 = \mathrm{soleil})) \\ -\mathrm{p}(\mathrm{rand}, \mathrm{q1} = \mathrm{nuage})\ log_2(p(\mathrm{rand}/q1 = \mathrm{nuage})) \\ -\mathrm{p}(\mathrm{pasrand}, \mathrm{q1} = \mathrm{nuage})\ log_2(p(\mathrm{pasrand}/q1 = \mathrm{nuage})) \end{matrix} = \begin{matrix} -3/8\ log_2(3/5) \\ -2/8\ log_2(2/5) \\ -1/8\ log_2(1/3) \\ -2/8\ log_2(2/3) \end{matrix}$$

Arbre de décision

En faisant ainsi pour chaque question, on trouve :

$$H(X/q_1) = 0.95$$
$$H(X/q_2) = 0.93$$
$$H(X/q_3) = 0.80$$
$$H(X/q_4) = 1$$

On choisit donc la question q3 qui minimise l'entropie: Humidité?
En répétant le processus,



Méthode générative classification bayésienne

Classification bayésienne

Il s'agit d'une **méthode générative**

On dispose d'un exemple $\boldsymbol{x}$ que l'on souhaite classer avec une étiquette *y.*
On calcule les densité de probabilité *a posteriori* de chaque classe:

$$p(y = k/\boldsymbol{x}) = \frac{p(\boldsymbol{x}/y = k)p(y = k)}{p(\boldsymbol{x})} = \frac{p(\boldsymbol{x}/y = k)p(y = k)}{\sum_{k'} p(\boldsymbol{x}/y = k')p(y = k')}$$

- $p(y = k)$: probabilité *a priori* de la classe k (avant d'observer $\boldsymbol{x}$)
- $p(\boldsymbol{x}/y = k)$: vraisemblance des observations
- $p(\boldsymbol{x})$: constante de normalisation
- $p(y = k/\boldsymbol{x})$: probabilité *a posteriori*

Apprentissage : estimer$p(\boldsymbol{x}/y = k)$et p$(y = k)$ sur la base d'apprentissage

Classification :

$$k = \underset{k}{\operatorname{argmax}}\, p(y = k/\boldsymbol{x}) = \underset{k}{\operatorname{argmax}} \frac{p(\boldsymbol{x}/y=k)p(y=k)}{p(\boldsymbol{x})} = \underset{k}{\operatorname{argmax}}\, p(\boldsymbol{x}/y = k)p(y = k)$$

Règle du Maximum A Posteriori (MAP)

Classification bayésienne

Reprenons l'exemple des truites et des saumons.

On dispose d'un ensemble d'exemples étiquetés (x_i, y_i)

- x_i: taille et la teinte de chaque poisson (dimension 2)
- $y_i \in \{truite, saumon\}$: classe

Commençons par les **probabilités *a priori*** :

p(truite)=$\frac{\text{Nombre de truites}}{\text{Nombre de poissons}} = \frac{200}{300} = \frac{2}{3}$

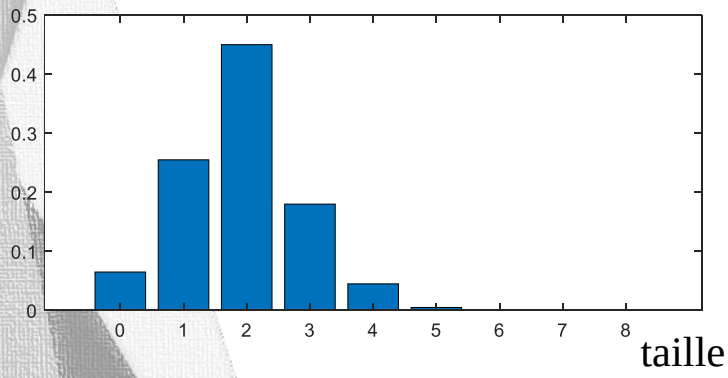
p(saumon)=$\frac{\text{Nombre de saumons}}{\text{Nombre de poissones}} = \frac{100}{300} = \frac{1}{3}$

Classification bayésienne

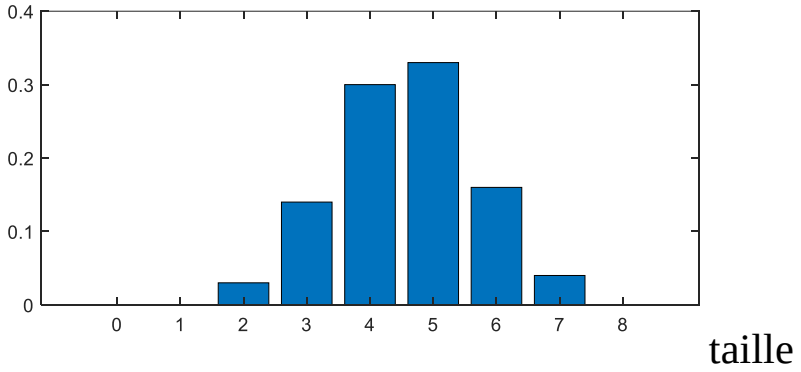
Il faut ensuite estimer **la vraisemblance des observations** à partir de la base d'apprentissage $p(\boldsymbol{x}/truite)$ et $p(\boldsymbol{x}/saumon)$

Travaillons dans un premier temps en **une seule dimension** en considérant uniquement la taille des poissons. Il faut estimer $p(taille/truite)$ et $p(taille/saumon)$. On peut, pour cela, calculer un histogramme normalisé (qui somme à 1) :

$p(x = taille/truite)$



$p(x = taille/saumon)$



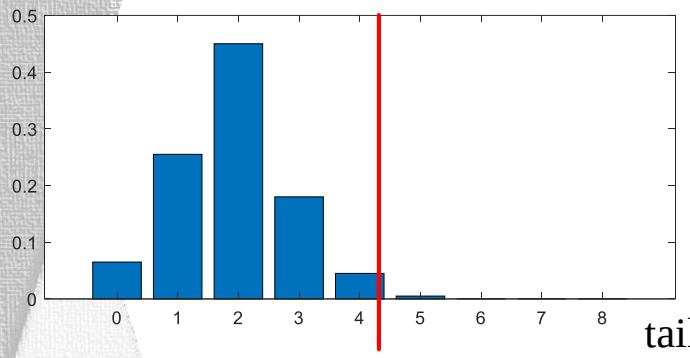
Classification bayésienne

Classification : **classer un poisson qui a une taille de 4 cm**

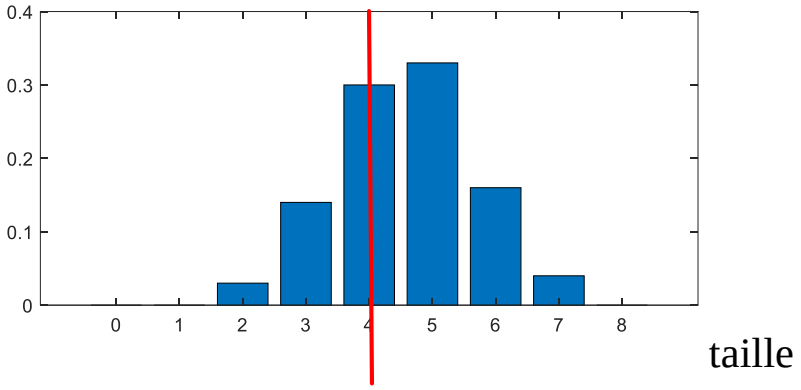
$p(x = 4/truite) = 0.045$

$p(x = 4/saumon) = 0.3$

$p(x = taille/truite)$



$p(x = taille/saumon)$



$$k = \underset{k}{\operatorname{argmax}}\, p(y = k/\boldsymbol{x}) = \underset{k}{\operatorname{argmax}}\, p(\boldsymbol{x}/y = k)p(y = k)$$

$$p(\mathbf{4}|truite)P(truite) = 0.045\text{x}\frac{2}{3} = 0.03$$

Saumon

$$p(\mathbf{4}|saumon)P(saumon) = 0.3\text{x}\frac{1}{3} = 0.1$$

Remarque : si on avait calculé les probabilités a posteriori,

$$p(truite|x) = \frac{p(\boldsymbol{x}|truite)P(truite)}{p(\boldsymbol{x}|truite)P(truite) + p(\boldsymbol{x}|saumon)P(saumon)} = \frac{0.045\text{x}\frac{2}{3}}{0.045\text{x}\frac{2}{3} + 0.3\text{x}\frac{1}{3}} = 0{,}03$$

$$p(saumon|x) = \frac{p(\boldsymbol{x}|saumon)P(saumon)}{p(\boldsymbol{x}|truite)P(truite) + p(\boldsymbol{x}|saumon)P(saumon)} = \frac{0.3\text{x}\frac{1}{3}}{0.045\text{x}\frac{2}{3} + 0.3\text{x}\frac{1}{3}} = 0{,}97$$

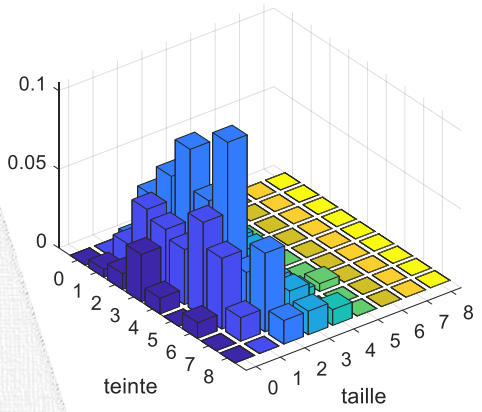
Les probabilités somment à 1

Classification bayésienne

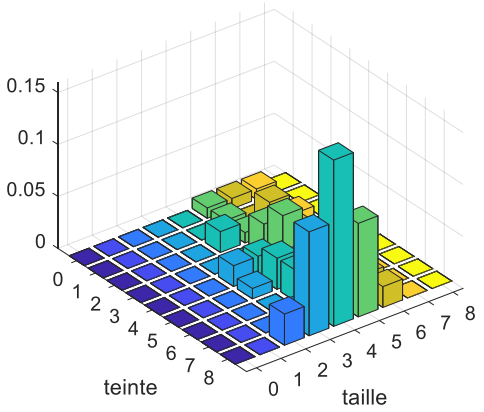
On va maintenant **utiliser les deux dimensions**. Chaque poisson est représenté par sa taille et sa teinte.
Les **probabilité *a priori*** ne changent pas : p(truite)=2/3 et p(saumon)=1/3

Pour la **vraisemblance des observations**, il faut maintenant passer en 2D :

$p(\boldsymbol{x}/truite)$



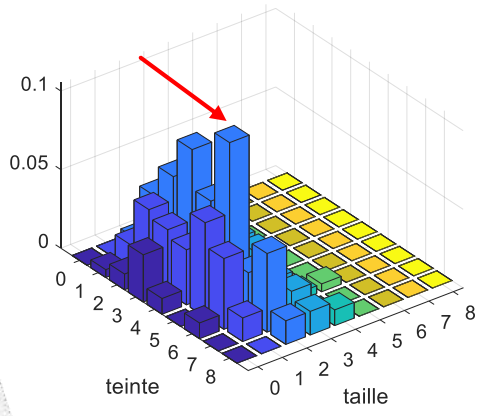
$p(\boldsymbol{x}/saumon)$



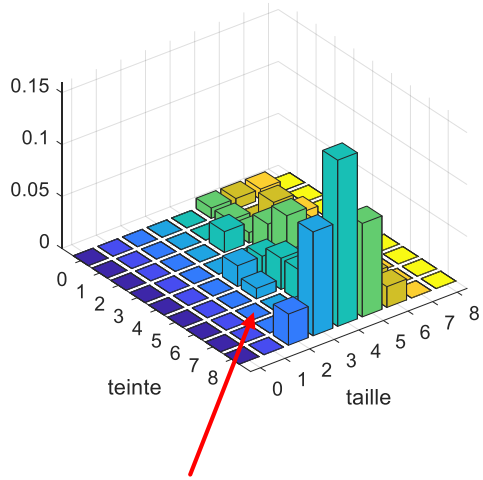
Classification bayésienne

Classification : classer un poisson qui a une **taille de 2 cm et une teinte de 6**

$p(\boldsymbol{x}/truite)$



$p(\boldsymbol{x}/saumon)$



$$p\left(\boldsymbol{x} = \begin{pmatrix}\boldsymbol{2}\\ \boldsymbol{6}\end{pmatrix} /truite\right) P(truite) = 0.105\text{x}\frac{2}{3} = 0.07$$

$$p\left(\boldsymbol{x} = \begin{pmatrix}\boldsymbol{2}\\ \boldsymbol{6}\end{pmatrix} /saumon\right) P(saumon) = 0\text{x}\frac{1}{3} = 0$$

Truite

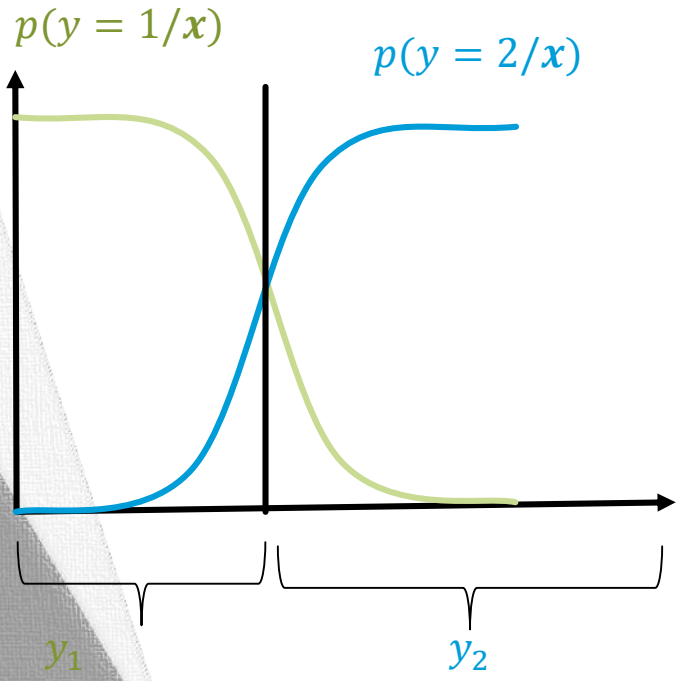
On en déduit donc que ce poisson est une truite

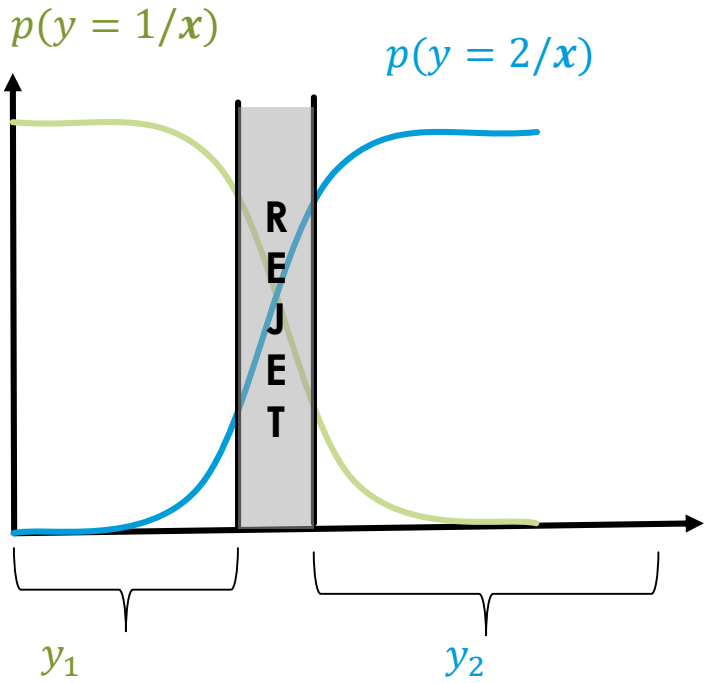
Classification bayésienne

Rq1 :

Décision avec rejet. Nous pouvons rejeter :

- les exemples tq la valeur maximale de $p(y|\boldsymbol{x})$<seuil
- les exemples qui ont leur deux plus grandes probabilités a posteriori similaires





Classification bayésienne

Rq 2 :

On a estimé la vraisemblance des observations avec des histogrammes, mais est ce robuste ?

- Comment fixer le pas de discrétisation ?
- Que se passe-t-il en grande dimension ?
- Y a-t-il façon de faire autrement ?

Estimation des densités de probabilité

Estimation des densités de probabilité

On se passe dans un cadre plus général d'estimation de densité de probabilité.

Connaissant un ensemble de N échantillons $\{\boldsymbol{x_i}\}_{\boldsymbol{i=1,\ldots,N}}$ de dimension *n* générés selon la loi de probabilité $p(\boldsymbol{x})$, comment estimer la densité de probabilité $p(\boldsymbol{x})$ à partir des N échantillons ?

Il existe deux grands types d'approches :

- les méthodes non paramétriques
- les méthodes paramétriques (la loi est fixée *a priori* et on en recherche les paramètres)

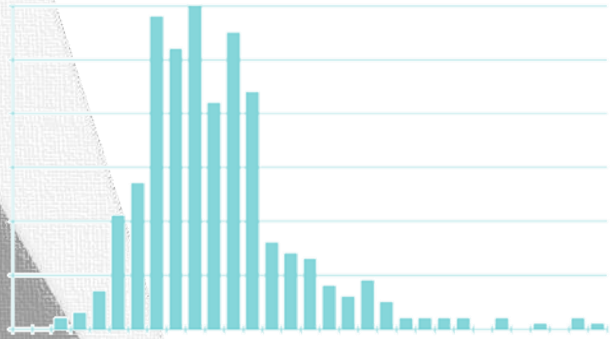
Rq : appliqué à la classification bayésienne, tous les exemples appartiennent à la même classe et on recherche la vraisemblance des observations pour cette classe $p(\boldsymbol{x}/y = k)$

Estimation des densités de probabilité

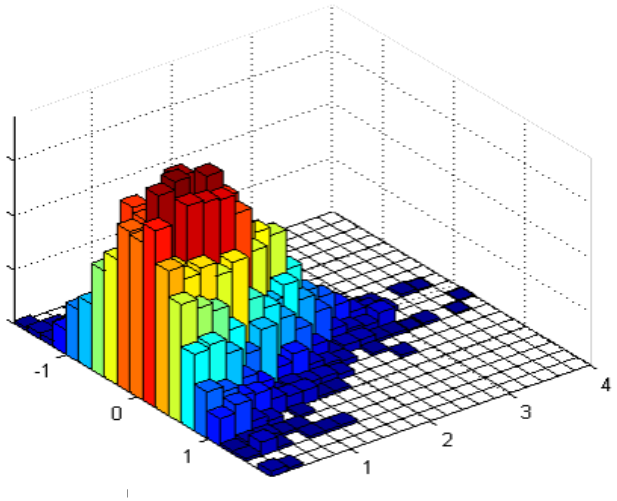
Histogramme (non paramétrique)

- On divise chaque dimension en cases (bins) de **largeur h**
- On compte le nombre d'échantillons x_i par case (divisé par N, le nombre d'échantillons)

En 1D



En 2D



Estimation des densités de probabilité

Plusieurs problèmes :

- **Problème de l'origine** : où la fixer
- **Problème du choix de h** (discrétisation)
- **Problème des grandes dimensions**

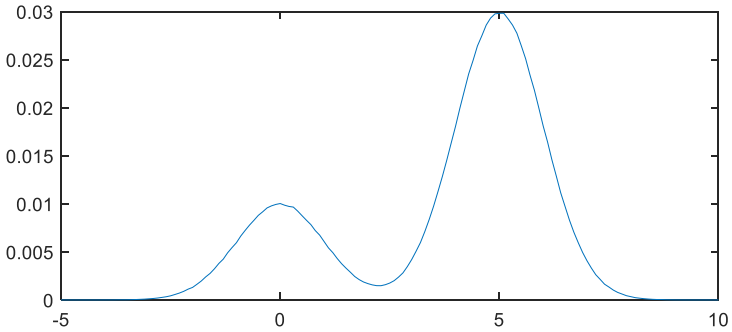
  Si les données sont de dimension 20 et que chaque dimension possède 5 cases, l'histogramme aura en tout $5^{20}=9.10^{13}$ cases

  ➔Il faudra une grosse base de données pour estimer $p(\boldsymbol{x})$

Estimation des densités de probabilité

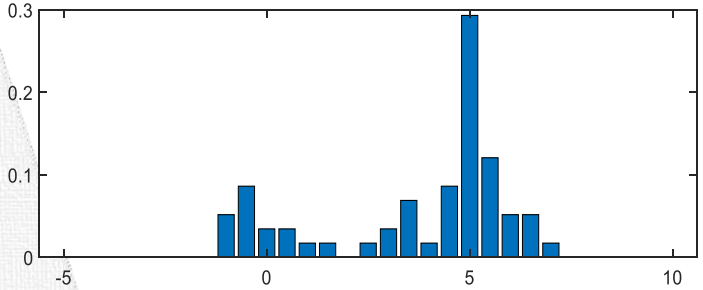
Problèmes liés au choix de *h*

$p(\boldsymbol{x})$

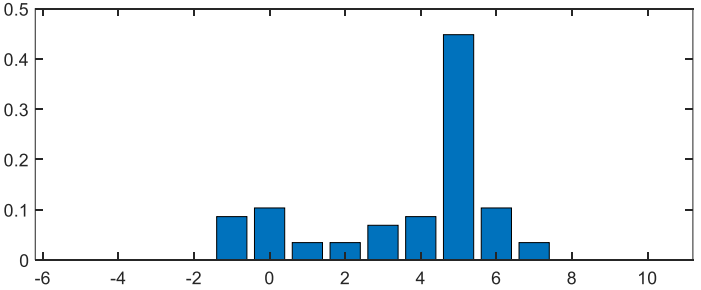


On tire *N* points selon *p(**x**)* puis on calcule l'histogramme normalisé (qui somme à 1) de ces points, en espérant retrouver $p(\boldsymbol{x})$

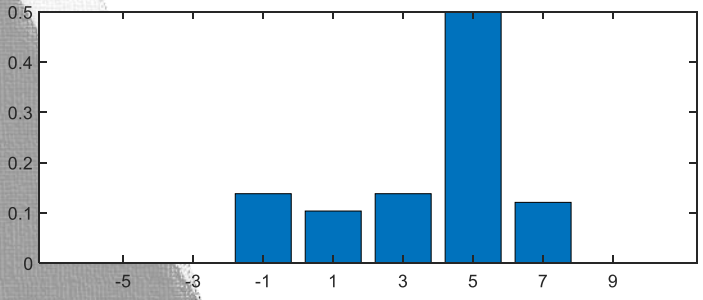
$\hat{p}(\boldsymbol{x})$ Pas de 0.5



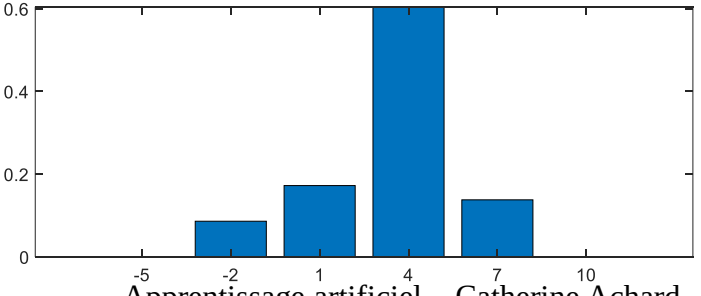
$\hat{p}(\boldsymbol{x})$ Pas de 1



$\hat{p}(\boldsymbol{x})$ Pas de 2



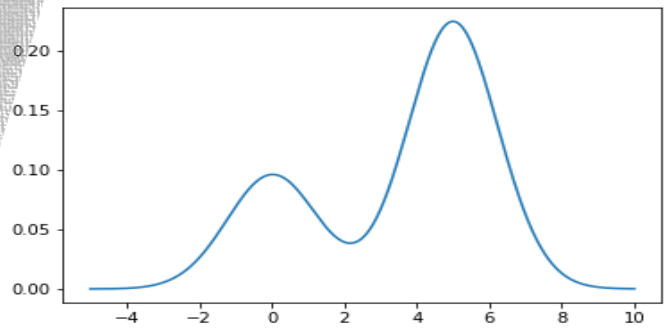
$\hat{p}(\boldsymbol{x})$ Pas de 3



Estimation des densités de probabilité

Problèmes liés au choix de l'origine

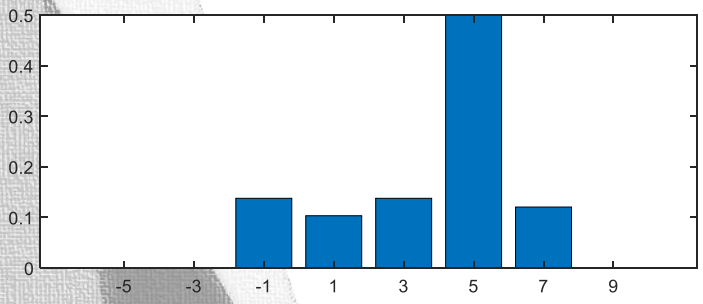
$p(\boldsymbol{x})$



On tire *N* points selon ***p(x)*** puis on calcule l'histogramme normalisé (qui somme à 1) de ces points, en espérant retrouver ***p(x)***

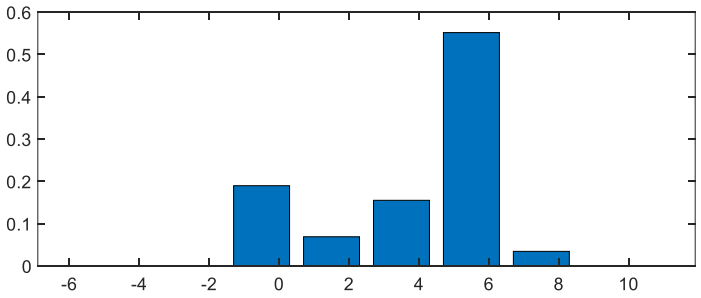
Pas de 2

$\hat{p}(\boldsymbol{x})$



Pas de 2, origine décalée de 0.5

$\hat{p}(\boldsymbol{x})$



Estimation des densités de probabilité

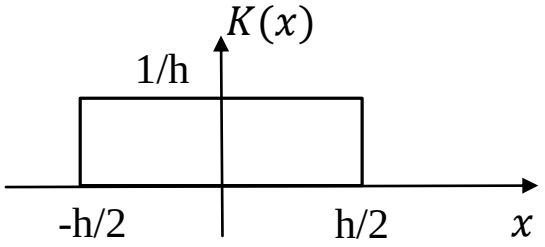
Estimation par noyau (Kernel density estimation)
Pour remédier au problème de l'origine,

$$\hat{p}(\boldsymbol{x}) = \frac{1}{N}\frac{\text{nombre d'échantillons dans } [\boldsymbol{x} - \boldsymbol{h}/2, \boldsymbol{x} + \boldsymbol{h}/2]}{h}$$

Pour trouver le nombre d'exemple qui tombent dans cet intervalle, on introduit la fenêtre :

$$K(x) = \begin{cases} 1/h \; si \; |h| < 1/2 \\ 0 \; sinon \end{cases}$$

K() est appelé **fenêtre de Parzen**



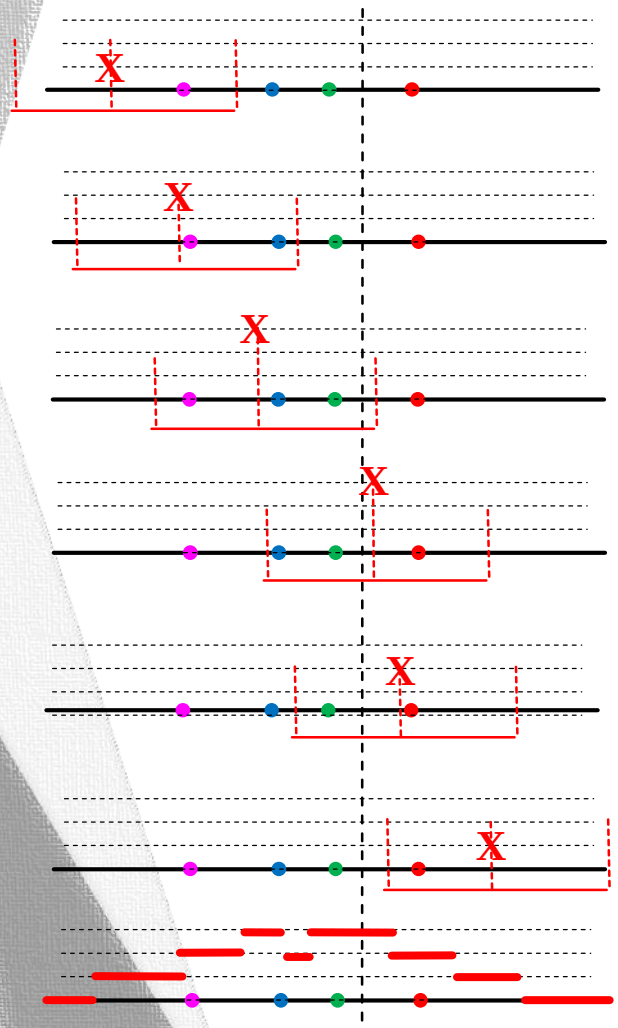
Le nombre de points qui tombe dans l'intervalle est défini par:

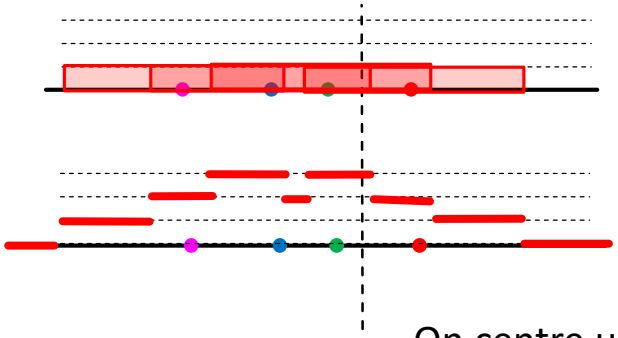
$$\hat{p}(\boldsymbol{x}) = \frac{1}{N}\sum_{i=1}^{N} K(x - \boldsymbol{x_i})$$

Cette équation s'étend à n'importe quelle dimension
Cette estimation est continue, elle est faite pour tout x

Estimation des densités de probabilité

Exemple avec 4 points :





On centre une fenêtre autour de chaque point et on fait la somme des fenêtres

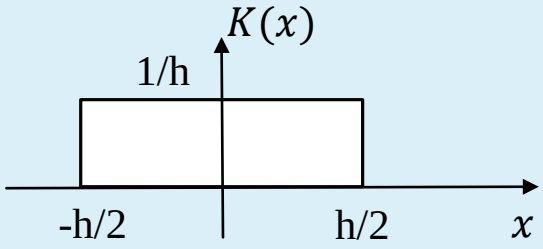
$$\hat{p}(\boldsymbol{x}) = \frac{1}{N}\sum_{i=1}^{N} K(x - \boldsymbol{x_i})$$

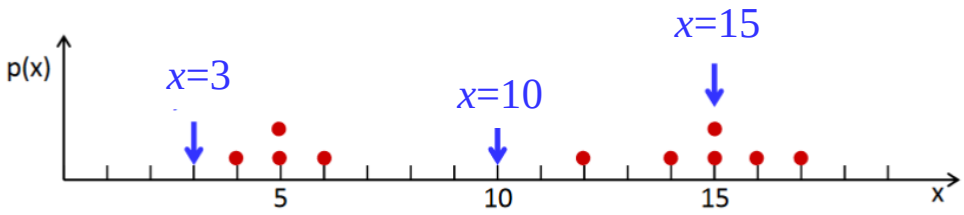
équivalent

$$\hat{p}(\boldsymbol{x}) = \frac{1}{N}\frac{\text{nombre d'échantillons dans } [\boldsymbol{x} - \boldsymbol{h/2}, \boldsymbol{x} + \boldsymbol{h/2}]}{h}$$

Estimation des densités de probabilité

Exercice
On donne un ensemble de points
X={4,5,5,6,12,14,15,15,16,17}
Estimer leur densité de probabilité $p(x)$ en x = 3, 10, 15 en utilisant les fenêtres de Parzen et h=3





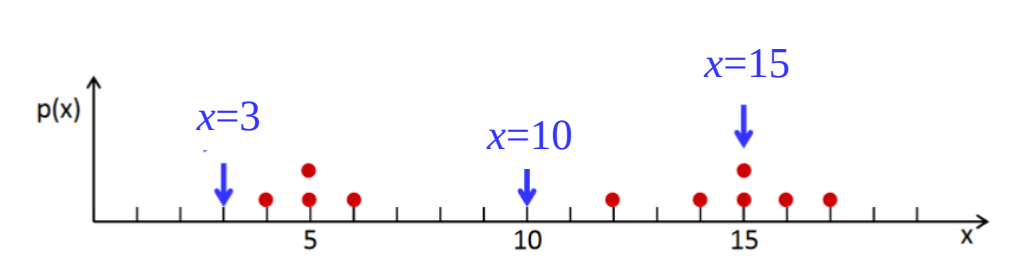
Estimation des densités de probabilité

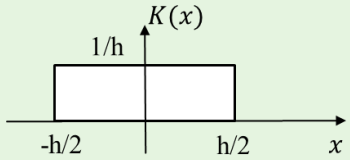
Exercice

On donne un ensemble de points

X={4,5,5,6,12,14,15,15,16,17}

Estimer leur densité de probabilité $p(x)$ en x = 3, 10, 15 en utilisant les fenêtres de Parzen et h=3





$$\hat{p}(x) = \frac{1}{N}\sum_{i=1}^{N} K(x - x_i) = \frac{1}{10}[K(x-4) + 2K(x-5) + \cdots]$$

Commençons en x=3:

$$\hat{p}(3) = \frac{1}{10}[1/3 + 2\text{x}0 + 0 + 0 + 0 + 2\text{x}0 + 0 + 0] = 0.0333$$

$$\hat{p}(10) = \frac{1}{10}[0 + 2\text{x}0 + 0 + 0 + 0 + 2\text{x}0 + 0 + 0] = 0$$

$$\hat{p}(15) = \frac{1}{10}[0 + 2\text{x}0 + 0 + 0 + 1/3 + 2\text{x}1/3 + 1/3 + 0] = 0.1333$$

Estimation des densités de probabilité

Inconvénients des fenêtres de Parzen :

- l'estimation de $p(x)$ est discontinue du fait de l'utilisation d'une fenêtre rectangulaire
- Tous les points à l'intérieur de la fenêtre sont pris en compte indépendamment de leur éloignement au point central

➔ **On préfère utiliser d'autres fenêtres nommées noyaux**

Rq : tous les noyaux doivent être tels que :

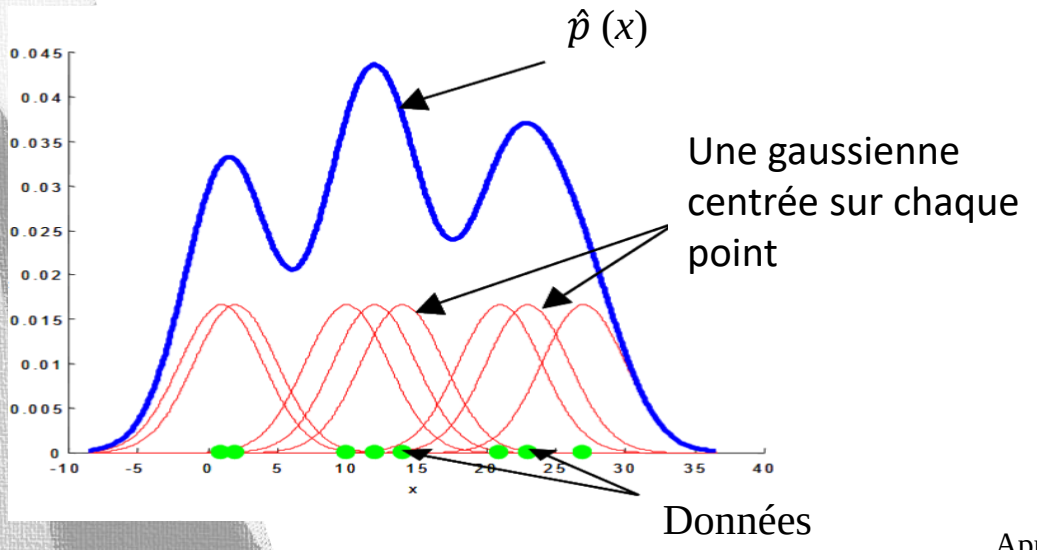
$$\int_{-\infty}^{\infty} K(x)dx = 1$$

Estimation par noyau (Kernel density estimation)

Un des noyaux le plus utilisé est le **noyaux gaussien :**

$$\mathrm{K}(\boldsymbol{x}) = \frac{1}{\sqrt{2\pi}h} exp\left(-\frac{x^2}{2h^2}\right)$$

Ceci revient à placer une gaussienne autour de chaque point et à sommer leur contribution

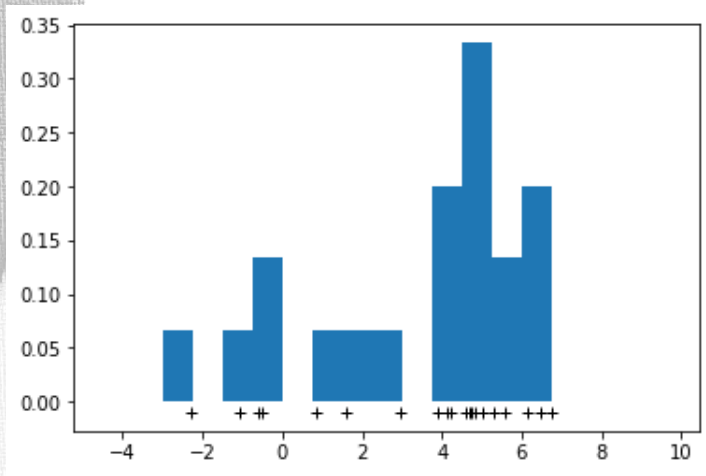


Estimation des densités de probabilité

Exemple avec h=0.75 sur 20 exemples

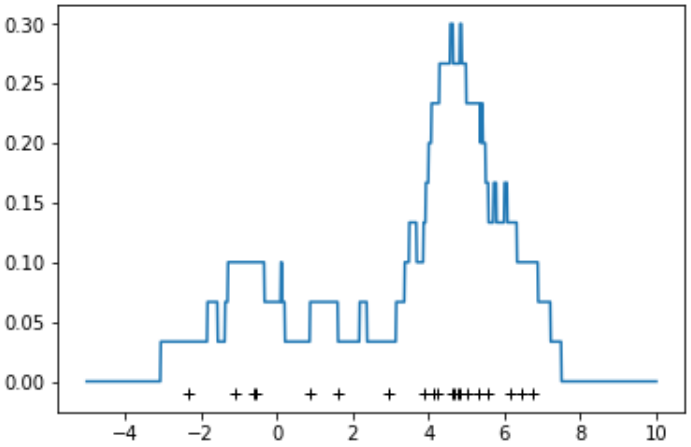
Histogramme normalisé

$\hat{p}(x)$



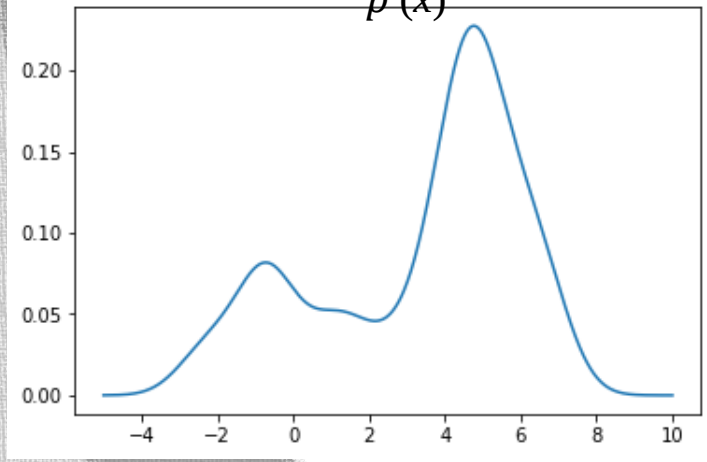
KDE et fenêtres de Parzen

$\hat{p}(x)$

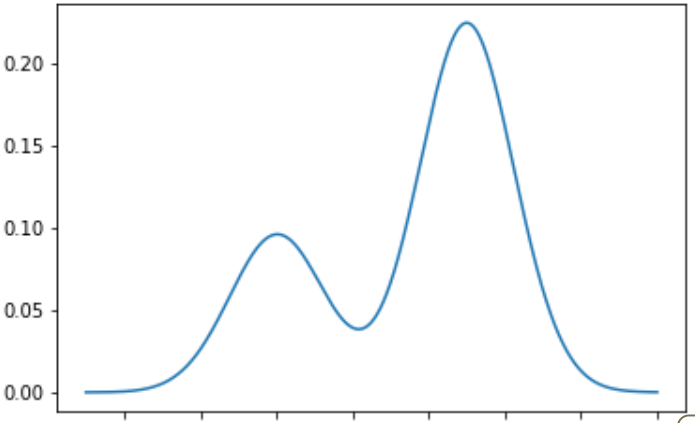


KDE et noyaux gaussiens

$\hat{p}(x)$



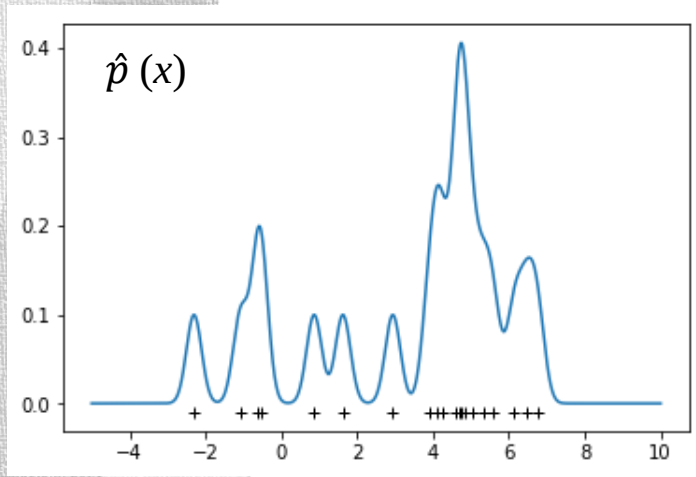
p(x) réel



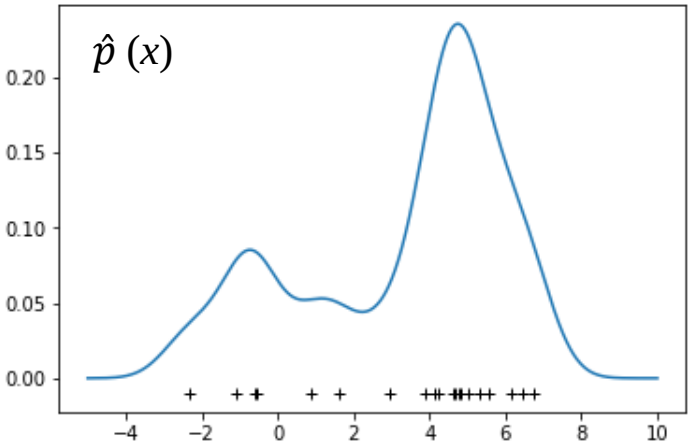
Estimation des densités de probabilité

Exemple avec 20 exemples, un noyau gaussien, en faisant varier h

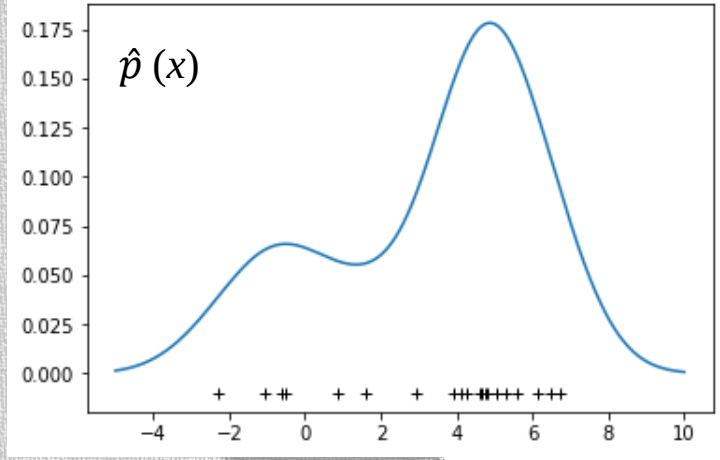
h=0.2



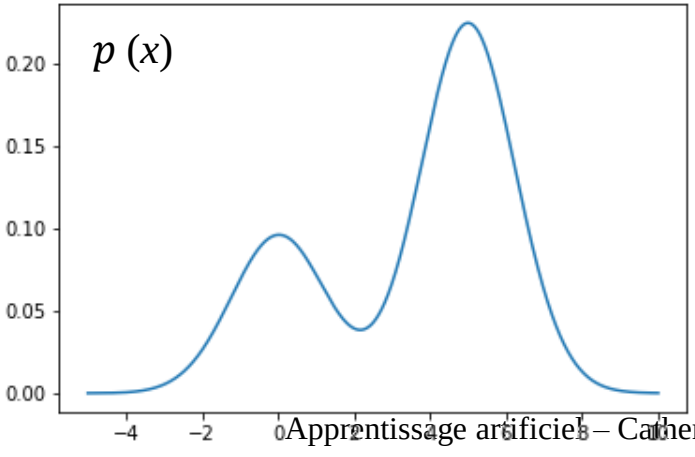
h=0.7



h=1.7



p(x) réel



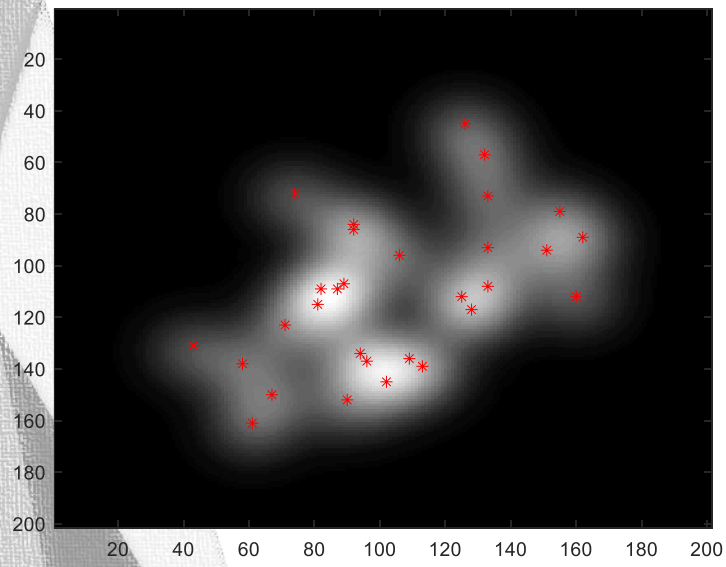
Estimation des densités de probabilité

Estimation par noyau (Kernel density estimation)

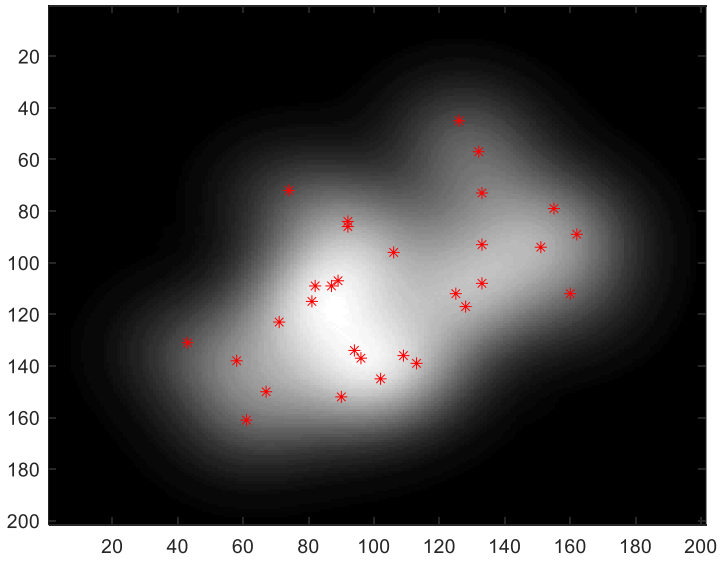
En 2D

La densité de probabilité

Noyau gaussien, h=0.02



h=0.05

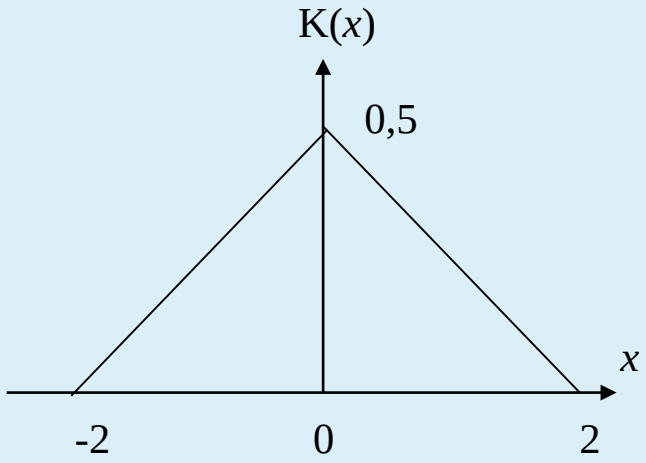


Estimation des densités de probabilité

On dispose des notes de mathématiques au bac de 10 élèves pour un certain lycée.
19 16 16 14 18 17 16 18 14 13

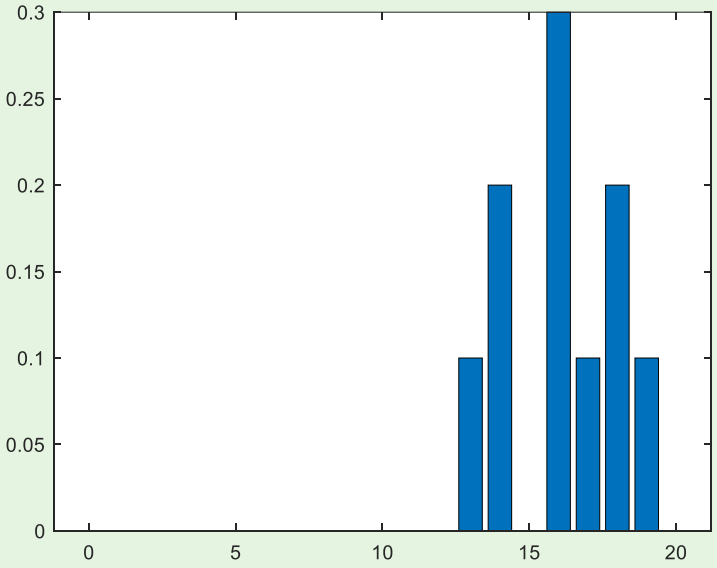
Estimer la densité de probabilité des notes avec:

- Un histogramme 1D discrétisé avec un pas de 1, d'origine -0,5
- La méthode des noyaux avec le noyau suivant :



Estimation des densités de probabilité

Avec l'histogramme



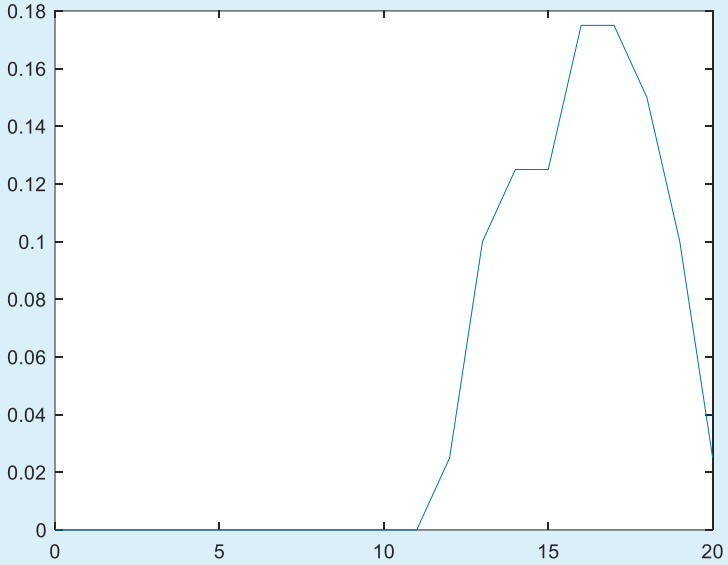
Pour l'histogramme, il suffit de compter le nombre d'occurrence de chaque note et de diviser par le nombre d'exemples.
Par exemple :

$$\hat{p}(19) = 1/10$$
$$\hat{p}(16) = 3/10$$

Estimation des densités de probabilité

Avec les noyaux triangulaires

19 16 16 14 18 17 16 18 14 13



$$\hat{p}(x) = \frac{1}{N}\sum_{i=1}^{N} K(x - x_i) = \frac{1}{10}\sum_{i=1}^{N} K(x - x_i)$$

Par exemple, pour x=15

$$\hat{p}(15) = \frac{1}{10}[K(15-13) + 2K(15-14) + 3K(15-16) + K(15-17) + 2K(15-18) + K(15-19)]$$

$$\hat{p}(15) = \frac{1}{10}[\quad 0 \quad + 2 * 0.125 \quad + 3 * 0.125 \quad + \quad 0 \quad + 2 * 0 \quad + 0 \quad]$$

$$\boldsymbol{\hat{p}(15) = 0.0625}$$

Estimation des densités de probabilité

Que faire quand on a peu d'échantillons de dimension élevée ?

La difficulté du problème est réduite si on connait *a priori* une forme paramétrique de la loi. Dans ce cas, il n'y a plus qu'à estimer les paramètres de la loi.

Ce problème est soluble par **l'estimation du maximum de vraisemblance**.

Estimation des densités de probabilité

Estimation du maximum de vraisemblance

Nous disposons de N échantillons $\boldsymbol{x}_i$ de dimension *n* tirés à partir de la loi $p(\boldsymbol{x})$.

$p(\boldsymbol{x})$ est modélisée par une loi paramétrique $\hat{p}(\boldsymbol{x}, \boldsymbol{\theta})$ de paramètres $\boldsymbol{\theta}$ que l'on cherche à estimer à partir des échantillons $\boldsymbol{x}_i$

Vraisemblance des observations :

$$L(\boldsymbol{\theta}/\boldsymbol{x}) = \prod_{i=1}^{N} \hat{p}(\boldsymbol{x_i}, \boldsymbol{\theta})$$

Il est souvent plus simple de travailler avec la log-vraisemblance:

$$l(\boldsymbol{\theta}/\boldsymbol{x}) = ln\left(\prod_{i=1}^{N} \hat{p}(\boldsymbol{x_i}, \boldsymbol{\theta})\right) = \sum_{i=1}^{N} \ln(\hat{p}(\boldsymbol{x_i}, \boldsymbol{\theta}))$$

On recherche $\boldsymbol{\theta}$ tq : $\widehat{\boldsymbol{\theta}} = \max_{\theta} L(\boldsymbol{\theta}/\boldsymbol{x})$

Si $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)^T$ est un vecteur de dimension p et que $\nabla_{\boldsymbol{\theta}} = \left(\frac{\partial}{\partial \theta_1}, \ldots, \frac{\partial}{\partial \theta_p}\right)^T$ est l'opérateur gradient, l'estimation des paramètres optimaux θ est telle que:

$\nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}/\mathbf{x}) = 0$ (p équations pour p inconnues)

Estimation des densités de probabilité

Exercice

Retrouver les paramètres d'une gaussienne 1D avec la règle du maximum de vraisemblance

Estimation des densités de probabilité

Exercice

Retrouver les paramètres d'une gaussienne 1D avec la règle du maximum de vraisemblance

➔$\hat{p}(\boldsymbol{x}, \boldsymbol{\theta})$ est une gaussienne de paramètres $\boldsymbol{\theta}=(\mu, \sigma)$:

$$\hat{p}(\boldsymbol{x}, \boldsymbol{\theta}) = \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

La vraisemblance est :

$$L(\theta/x) = \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

$$\mathrm{l}(\theta/x) = \sum_{i=1}^{N} \left[-\ln(\sigma\sqrt{2\pi}) - \frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

$$\mathrm{l}(\theta/x) = -\sum_{i=1}^{N} \ln(\sigma) \; - \sum_{i=1}^{N} \ln(\sqrt{2\pi}) \; - \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma^2}$$

Estimation des densités de probabilité

Exercice

Retrouver les paramètres d'une gaussienne 1D avec la règle du maximum de vraisemblance

$$l(\theta/x) = -\sum_{i=1}^{N} \ln(\sigma) \; - \sum_{i=1}^{N} \ln(\sqrt{2\pi}) \; - \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma^2}$$

$$\frac{\partial l(\theta/x)}{\partial \sigma} = -\sum_{i=1}^{N} \frac{1}{\sigma} + \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{\sigma^3} = 0$$
$$-\frac{N}{\sigma} + \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{\sigma^3} = 0$$
$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

$$\frac{\partial l(\theta/x)}{\partial \mu} = \sum_{i=1}^{N} \frac{2(x_i-\mu)}{2\sigma^2} = 0$$
$$\sum_{i=1}^{N}(x_i - \mu) = 0$$
$$\sum_{i=1}^{N} x_i - N\mu = 0$$
$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Estimation des densités de probabilité

Loi de Bernouilli (estimation paramétrique)

Si x est une variable binaire, alors

$$Bern(x = 1) = \mu \quad \text{et} \quad Bern(x = 0) = 1 - \mu$$

Ou encore

$$Bern(x) = \mu^x (1 - \mu)^{1-x}$$

On montre que :

$$\mathbb{E}[x] = \mu \quad \text{et} \quad var[x] = \mu(1 - \mu)$$

Et, avec l'estimation du maximum de vraisemblance:

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

Ces résultats peuvent être retrouvés par le calcul

Loi binomiale (estimation paramétrique)

Supposons que l'on tire N échantillons binaires selon la loi de Bernouilli. La variable aléatoire x qui compte le nombre de réalisations de 1 parmi ces N échantillons suit une loi binomiale de paramètres N et λ.
x peut donc prendre toutes les valeurs entières de 0 à N et

$$p(x) = \frac{N!}{(N-x)!\,x!}\lambda^x(1-\lambda)^{\mathrm{N}-x}$$

On montre alors que :

$$\mathbb{E}[x] = \mathrm{N}\lambda \quad \text{et} \quad var[x] = \sqrt{N\lambda(1-\lambda)}$$

Ces résultats peuvent être retrouvés par le calcul

Estimation des densités de probabilité

Loi uniforme (estimation paramétrique)

La variable aléatoire continue x suit une loi uniforme sur l'intervalle [a,b] si:

$$p(x) = \frac{1}{b-a}$$

On a alors

$$\mathbb{E}[x] = \frac{b-a}{2} \quad \text{et} \quad var[x] = \frac{(b-a)^2}{12}$$

Estimation des densités de probabilité

Loi normale mono variable (estimation paramétrique)

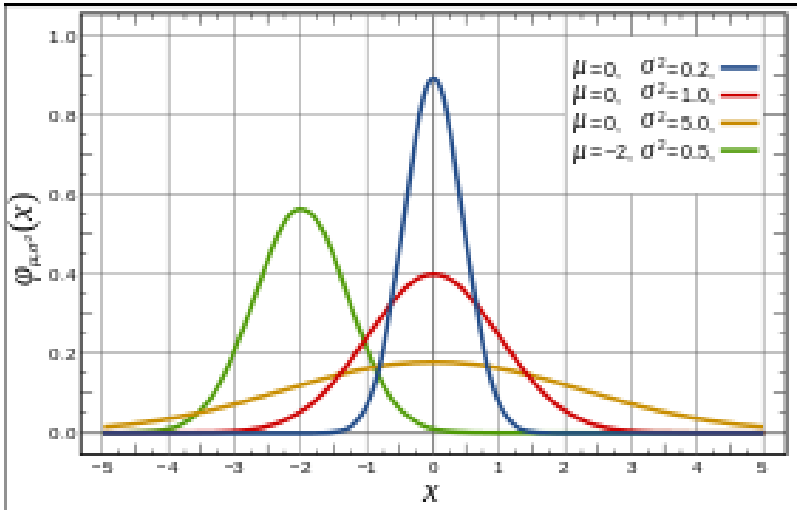
Elle est définie par :

$$p(x) = \mathcal{N}(x/\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Et : $\mathbb{E}[x] = \mu$ et $var[x] = \sigma^2$

L'estimation du maximum de vraisemblance conduit à:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \quad \text{et} \quad \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$



Problème

Problème

Beaucoup de fraudes apparaissent dans les animaleries sur les lapins : alors que les gens pensent acheter un lapin nain, on leur vend un lapin normal qui devient vitre très imposant. Le problème est que bébé, les deux espèces se ressemblent énormément. Aussi, deux mesures ont été réalisées à partir de prélèvements sanguins sur 20 lapins de chaque espèce. Les résultats sont présentés dans le tableau 1. Ces mesures sont reportées graphiquement figure 1.

Afin de comparer différentes méthodes de reconnaissance des formes, on dispose d'une base de test composée de 5 lapins de chaque espèce (tableau 2)

On considère que les 2 classes sont équiprobables.

1. Réaliser la classification des exemples de test avec l'algorithme des 1PPV. Donner la matrice de confusion

Base de références

Lapins nains		Lapins normaux	
Mesure 1	Mesure 2	Mesure 1	Mesure 2
153.7	125.4	157.7	141.5
145.6	105.5	142.2	136.9
108.3	110.3	167.3	134.9
124.4	122.0	137.0	147.1
135.5	116.5	145.6	153.3
127.6	132.9	155.8	161.2
133.9	133.4	167.2	140.5
117.7	114.1	143.0	141.6
119.2	128.7	132.0	133.6
121.7	133.9	148.0	156.1
104.2	123.2	126.8	139.2
148.4	136.2	137.2	135.2
120.1	130.6	142.7	161.8
133.7	122.1	154.8	148.0
102.9	128.7	154.3	147.1
98.4	121.9	139.7	129.9
118.1	115.8	149.2	144.3
114.9	123.5	136.7	145.2
143.8	120.3	146.6	129.0
132.1	116.3	159.1	137.7
Moy 125.2	**Moy 123.1**	**Moy 147.1**	**Moy 143.2**
Std 15.6	**Std 8.3**	**Std 11.1**	**Std 9.5**

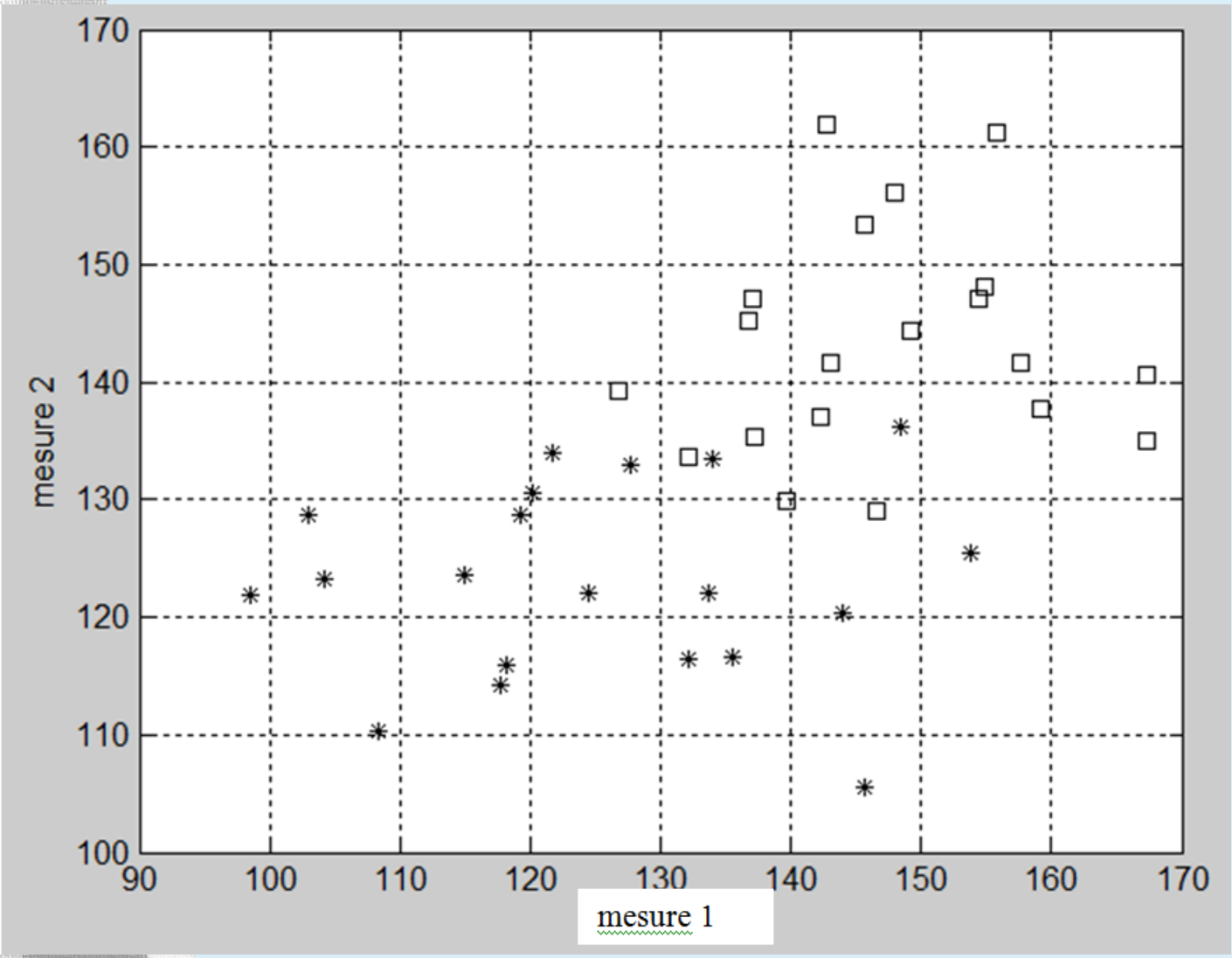
Tableau 1

Base de test

Lapins nains		Lapins normaux	
Mesure 1	Mesure 2	Mesure 1	Mesure 2
135	115	137	152
125	125	155	152
155	125	132	142
135	132	142	135
129	139	147	136

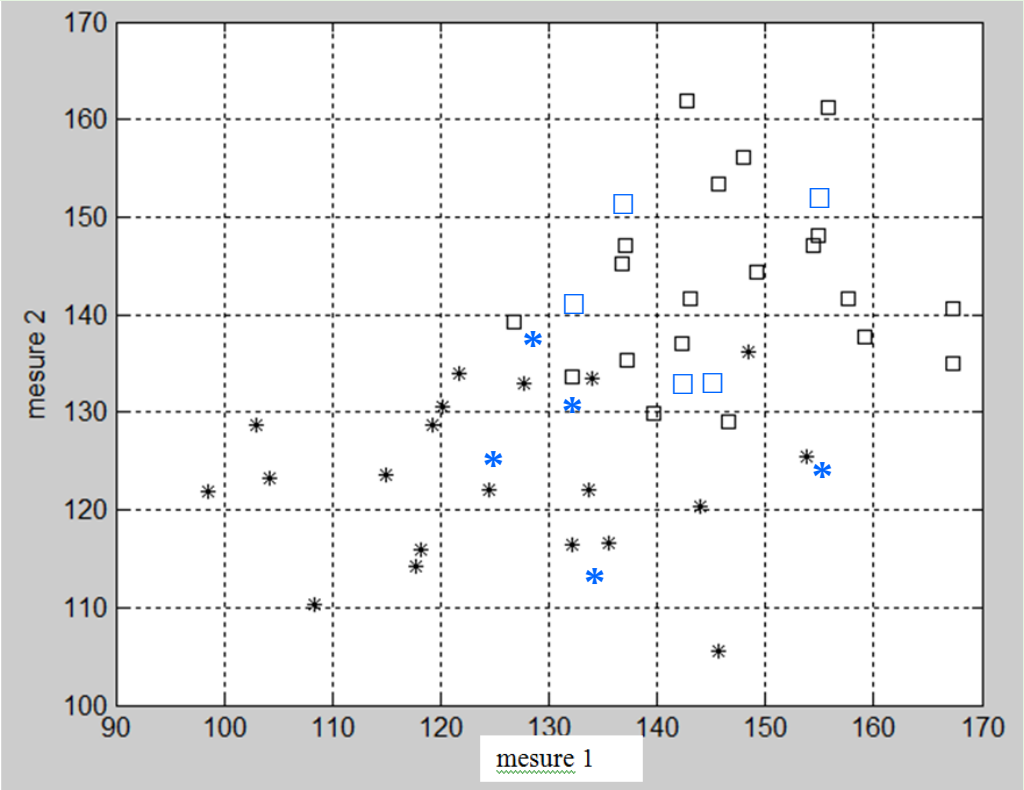
Tableau 2

Problème


170
160
150
140
130
120
110
100
90
100
110
120
130
140
150
160
170
mesure 2
mesure 1

Problème

Question 1



Algorithme du 1PPV

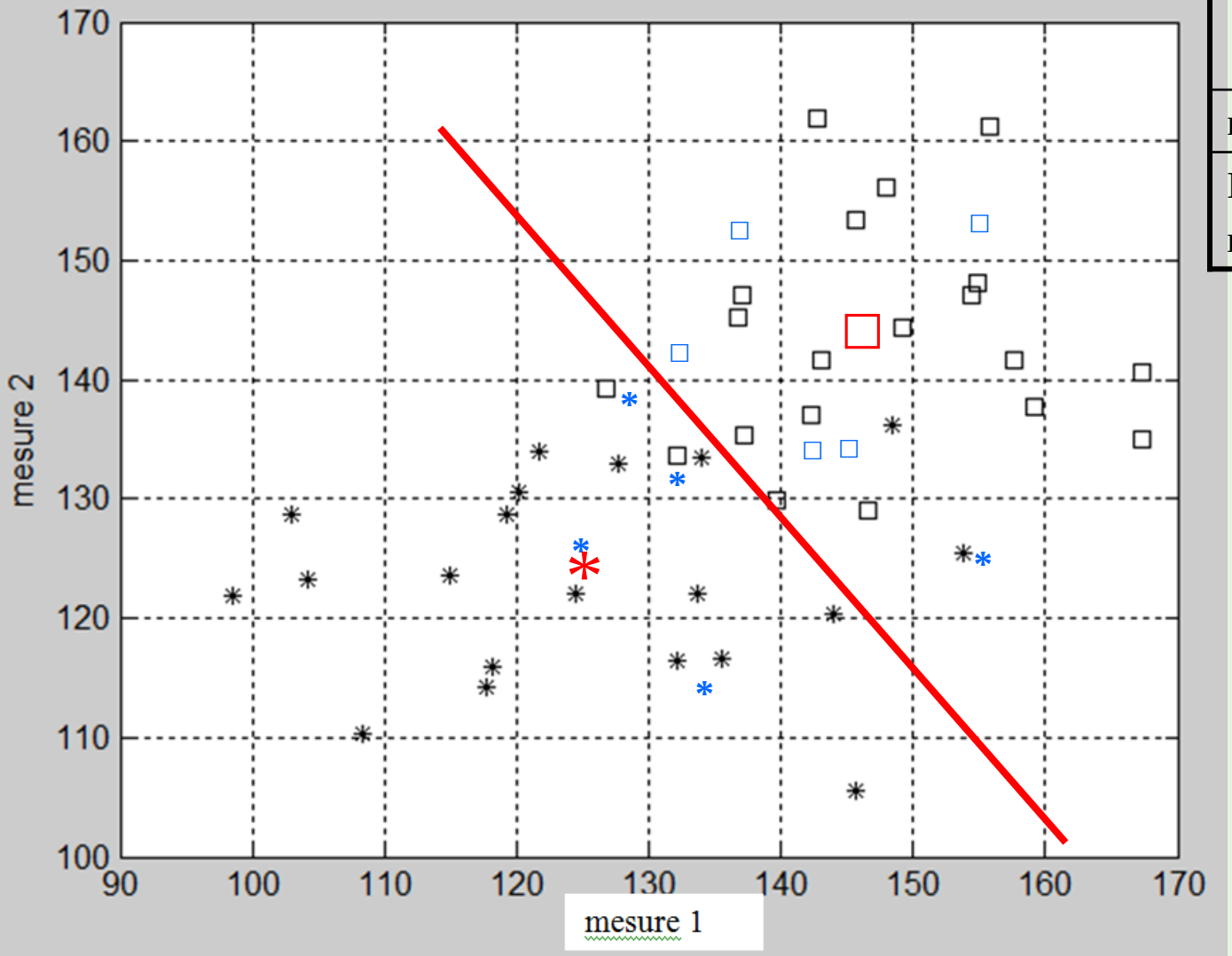
	nain	normaux
nain	3	2
normaux	0	5

Problème

2. On décide de modéliser chaque classe uniquement par sa **moyenne** et de classifier avec l'approche **nearest mean**. Réaliser la classification et donner le taux de reconnaissance

Problème

Question 2



	nain	Normaux
nain	4	1
Normaux	0	5

Taux de reconnaissance

$$= \frac{9}{10} = 90\%$$

Problème

3. On souhaite classer ces exemples avec une classification bayésienne. Pour cela, on modélise la vraisemblance des observations $p(x/y)$ de chaque classe par des histogrammes bi-dimensionnel. Remplir les cases qui nous sont utiles des tableaux ci-dessous.
Réaliser ensuite la classification (en cas d'égalité, l'exemple sera rejeté). Donner la matrice de confusion. En déduire le taux de reconnaissance et le taux de rejet.

Lapins nains							
m2 / m1	100-110	110-120	120-130	130-140	140-150	150-160	160-170
90-100							
100-110							
110-120							
120-130							
130-140							
140-150							
150-160							
160-170							

Lapins normaux							
m2 / m1	100-110	110-120	120-130	130-140	140-150	150-160	160-170
90-100							
100-110							
110-120							
120-130							
130-140							
140-150							
150-160							
160-170							

4. Reprendre les mêmes questions avec les nouveaux tableaux. Conclusion.

Lapins nains				
	90-110	110-130	130-150	150-170
90-110				
110-130				
130-150				
150-170				

Lapins normaux				
	90-110	110-130	130-150	150-170
90-110				
110-130				
130-150				
150-170				

Problème

Question 3

$p(x/nain) * 20$

Lapins nains							
m2 / m1	100-110	110-120	120-130	130-140	140-150	150-160	160-170
90-100	0	0	1	0	0	0	0
100-110	0	1	2	0	0	0	0
110-120	0	2	2	0	0	0	0
120-130	0	0	1	3	0	0	0
130-140	0	2	1	1	0	0	0
140-150	1	0	1	1	0	0	0
150-160	0	0	1	0	0	0	0
160-170	0	0	0	0	0	0	0

$p(x/normaux) * 20$

Lapins normaux							
m2 / m1	100-110	110-120	120-130	130-140	140-150	150-160	160-170
90-100	0	0	0	0	0	0	0
100-110	0	0	0	0	0	0	0
110-120	0	0	0	0	0	0	0
120-130	0	0	0	1	0	0	0
130-140	0	0	1	2	2	0	0
140-150	0	0	1	1	2	2	1
150-160	0	0	0	1	3	0	1
160-170	0	0	0	1	1	0	0

$$p(y|x) = \frac{p(\boldsymbol{x}|y)p(y)}{p(\boldsymbol{x})}$$

Comme les classes sont équiprobables, il suffit de comparer les cases des histogrammes

	nain	normaux
nain	4	1
normaux	0	1

Taux de reconnaissance = $\frac{5}{10} = 50\%$

Taux de rejet= $\frac{4}{10} = 40\%$

Problème

Question 4

$p(x/nain) * 20$

Lapins nains				
	90-110	110-130	130-150	150-170
90-110	0	3	0	0
110-130	0	5	3	0
130-150	1	4	2	0
150-170	0	1	0	0

$p(x/normaux) * 20$

Lapins normaux				
	90-110	110-130	130-150	150-170
90-110	0	0	0	0
110-130	0	0	1	0
130-150	0	2	7	3
150-170	0	0	6	1

	nain	normaux
nain	4	1
normaux	0	5

Taux de reconnaissance = $\frac{9}{10} = 90\%$

Taux de rejet= $\frac{0}{10} = 0\%$